6. *Correlation*

Correlation quantifies the extent to which two quantitative variables, X and Y, "go together." When high values of X are associated with high values of Y, a positive correlation exists. When high values of X are associated with low values of Y, a negative correlation exists.

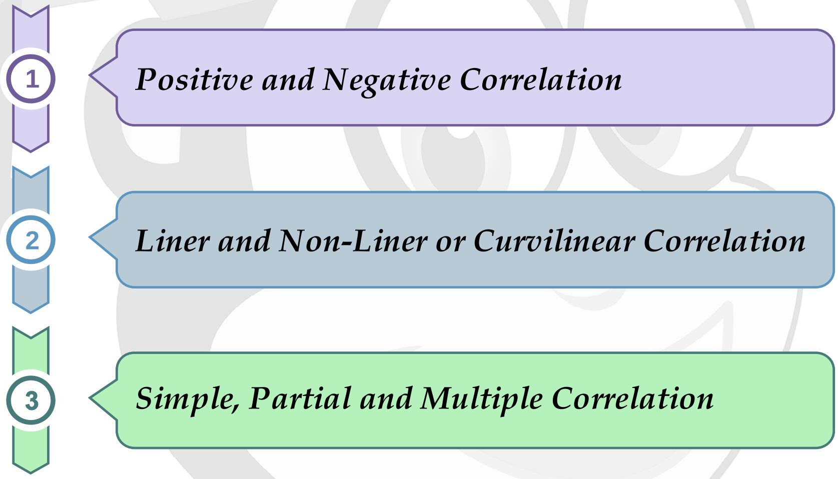
Correlation is a Powerful statistical tool that provides quantitative expression of the manner or extent to which events are related mathematically.

> **“ *According to Prof. King*** *"Correlation means that between two series or group of data these exists some casual connection "*

> **“ *According to Boddington*** *"Whenever some define connections exit between the two or more groups, classes or series of data there is said to be correlation."*

Thus correlation analysis helps us in determining the degree of relationship between two or more variables. It does not tell us about cause and effect relationship.

Types of Correlations



1 *Positive and Negative Correlation*

2 *Liner and Non-Liner or Curvilinear Correlation*

3 *Simple, Partial and Multiple Correlation*

(1) Positive and Negative Correlation

If change in two variables are in same direction i.e. increase in price and increase in supply, increase in father age and increase in son age etc.

If variations and fluctuations in two variable are in opposite direction to in other word increase in one variable is associated with corresponding decrease in others or vice versa. Then correlation is said to be negative. i.e. increases in price associated with decrease in demand and vice versa.

(2) Linear and Non Linear (Curvilinear) Correlation

The distinction between linear and non- linear correlation is based upon the constancy of the ratio of change between the two variables. If amount of change in one variable with the constant ratio of change in other variable the correlation is called liner correlation. For example, if in a factory of raw material or number of direct worker are doubled the production is also doubled and vice versa.

On the hand correlation would be called curvilinear (Non Liner) if the amount of change in one variable does not bear a constant ratio of change in another variable. For example, the amount spent on advertisement will not bring the change in the amount of sales in the same ratio.

(3) Simple, Partial and Multiple Correlation

When two variables are studied it is a case of simple correlation. On other hand when three or more variables are studied it is a problem of multiple or partial correlation.

When three or more variable are studied simultaneously it is called multiple correlation but when a study of yield correlation more than two variables are studied but consider the influence of third variable on the two variables it is called partial correlation.

Degree of Correlation

Degree	Positive	Negative
Perfect	1	– 1
High	+ 0.75 to + 1	– 0.75 to – 1
Moderate	+ 0.25 to + 0.75	– 0.25 to – 0.75
Low	+ 0 to + 0.25	– 0 to – 0.25
Absence	0	0

(i) Perfect Correlation

If there is any change in the value of one variable, the value of the other variable is changed in a fixed proportion. If the variations in two variables are in constant ratio in same direction the correlation is perfect positive. On other hand if correlation in two variables is in constant ration in opposite direction the correlation is perfect negative.

The correlation between them is said to be a perfect correlation. It is indicated numerically as +1 and –1.

(ii) Absence of Correlation

If variations in two variables are not corresponding to each other it is case of absence of correlation. It is denoted by 0.

Limited Degree of Correlation

If coefficient of correlation is more than zero but less than one it is called limited correlation.

- High degree correlation
- Moderate degree correlation
- Low degree Correlation

Methods of Correlation

Graphical Method	Mathematical Method
1 Scatter diagram method	1 Karl Pearson's
2 Simple Graphic Method	2 Spearman's
	3 Concurrent Deviation

(1) Karl Pearson's

Karl Pearson’s Coefficient of Correlation

Focus Formula

According to Karl Pearson’s method, the coefficient of correlation is measured as

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

Where,

r = Coefficient of Correlation

$x = X - \bar{X}$

$y = Y - \bar{Y}$

σ_x = Standard deviation of X series

σ_y = Standard deviation of Y series

N = Number of observations.

A Modified Version of Karl Pearson’s Formula

In it there is no need to calculate standard deviation of ‘X’ and ‘Y’. Coefficient of correlation may be worked out directly using the following formula :

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Here, $x = (X - \bar{X})$; $y = (Y - \bar{Y})$

Illustration : Calculate the coefficient of correlation between the age of husbands and wives :

Age of Husband (Years)	21	22	28	32	35	36
Age of Wife (Years)	18	20	25	30	31	32

Solution :

Age of Husband (X)	**Deviation** $(x = X - \bar{X})$ $\bar{X} = 29$	**Square of Deviation (x^2)**	**Age of Wife (Y)**	**Deviation** $(y = Y - \bar{Y})$ $\bar{Y} = 26$	**Square of Deviation (y^2)**	**Multiple of Deviation (xy)**
21	– 8	64	18	– 8	64	64
22	– 7	49	20	– 6	36	42
28	– 1	1	25	– 1	1	1
32	+ 3	9	30	+ 4	16	12
35	+ 6	36	31	+ 5	25	30
36	+ 7	49	32	+ 6	36	42
$\Sigma X = 174$	$\Sigma x = 0$	$\Sigma x^2 = 208$	$\Sigma Y = 156$	$\Sigma y = 0$	$\Sigma y^2 = 178$	$\Sigma xy = 191$

$$x = \left(X - \bar{Y}\right);\ y = \left(Y - \bar{Y}\right)$$

$$\overline{X} = \frac{\sum X}{N} = \frac{174}{6} = 29;\ \overline{Y} = \frac{\sum Y}{N} = \frac{156}{6} = 26$$

$\Sigma xy = 191,\ \Sigma x^2 = 208,\ \Sigma y^2 = 178$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

$$r = \frac{191}{\sqrt{208 \times 178}}$$

$$= \frac{191}{\sqrt{37{,}024}} = \frac{191}{192.42} = 0.993$$

Coefficient of Correlation (r) = 0.993.

Thus, there is a high degree of positive correlation between the age of husband and wife.

Short-Cut Method

This method is used when mean value is not in whole number but in fractions. In this method, deviation is calculated by taking the assumed mean of both the series. It involves the following steps :

(1) Any convenient value in X and Y series is taken as assumed mean A_X and A_Y.

(2) With the help of assumed mean of both the series, deviation of the values of individual variable, i.e., dx $(X - A_X)$ and dy $(Y - A_Y)$ are calculated.

(3) Σdx and Σdy are found by adding the deviations.

(4) Deviations of the two series are multiplied, as dx.dy, and the multiples added up to obtain Σdxdy.

(5) Squares of the deviations dx^2 and dy^2 are added up to find out Σdx^2 and Σdy^2.

(6) Finally, coefficient of correlation is calculated using the following formula :

Focus Formula

$$r = \frac{\Sigma\, dxdy - \dfrac{(\Sigma\, dx) \times (\Sigma\, dy)}{N}}{\sqrt{\Sigma\, dx^2 - \dfrac{(\Sigma\, dx)^2}{N}} \times \sqrt{\Sigma\, dy^2 - \dfrac{(\Sigma\, dy)^2}{N}}}$$

Here, dx = Deviation of X series from the assumed mean = (X – A)
dy = Deviation of Y series from the assumed mean = (Y – A)
Σdxdy = Sum of the multiple of dx and dy
Σdx^2 = Sum of square of dx
Σdy^2 = Sum of square of dy
Σdx = Sum of deviation of X series
Σdy = Sum of deviation of Y series
N = Total number of items.

Illustration : Calculate coefficient of correlation between the price and quantity supplied :

Price (₹)	4	6	8	15	20
Supply (kg)	10	15	20	25	30

Solution :

Price (X)	Deviation (dx = X – A) A = 8	Square of Deviation (dx^2)	Supply (Y)	Deviation (dy = Y – A) A = 20	Square of Deviation (dy^2)	Multiple of Deviations (dxdy)
4	– 4	16	10	– 10	100	40
6	– 2	4	15	– 5	25	10
8(A)	0	0	20(A)	0	0	0
15	7	49	25	5	25	35
20	12	144	30	10	100	120
N = 5	Σdx = 13	$\Sigma dx^2 = 213$	N = 5	Σdy = 0	$\Sigma dy^2 = 250$	Σdxdy = 205

$$r = \frac{\sum dxdy - \frac{\left(\sum dx\right)\times\left(\sum dy\right)}{N}}{\sqrt{\sum dx^2 - \frac{\left(\sum dx\right)^2}{N}} \times \sqrt{\sum dy^2 - \frac{\left(\sum dy\right)^2}{N}}}$$

$$r = \frac{205 - \frac{(13)\times(0)}{5}}{\sqrt{213 - \frac{(13)^2}{5}} \times \sqrt{250 - \frac{(0)^2}{5}}}$$

$$r = \frac{205 - 0}{\sqrt{213 - 33.80} \times \sqrt{250 - 0}}$$

$$r = \frac{205}{\sqrt{179.20} \times \sqrt{250}}$$

$$= \frac{205}{13.39 \times 15.81}$$

$$r = \frac{205}{211.70}$$

$$= +0.97$$

Coefficient of Correlation (r) = +0.97.

This is a situation of a high degree of positive correlation.

Step-Deviation Method

The method involves the following steps :

(1) Repeat Step-1 and Step-2 of the short-cut method.

(2) Now divide ‘dx’ and ‘dy’ by some common factor as dx′ = $\frac{dx}{C_1'}$ dy = $\frac{dy}{C_2}$; here C_1 is common factor for series X and C_2 is common factor for series Y. And dx′ and dy′ are step-deviations.

(3) Σdx′ and Σdy′ are found by adding the deviations.

(4) Deviations of the two series are multiplied, as dx′ × dy′, and the multiples added up to obtain Σdx′ dy′.

(5) Squares of the deviations dx'^2 and dy'^2 are added up to find out $\Sigma dx'^2$ and $\Sigma dy'^2$.

(6) Finally, coefficient of correlation is calculated using the following formula.

Focus Formula

$$r = \frac{\sum dx'dy' - \frac{(\sum dx')\times(\sum dy')}{N}}{\sqrt{\sum dx'^2 - \frac{(\sum dx')^2}{N}} \times \sqrt{\sum dy'^2 - \frac{(\sum dy')^2}{N}}}$$

Illustration : Calculate coefficient of correlation between the price and quantity demanded :

Price (₹)	5	10	15	20	25
Demand (kg)	40	35	30	25	20

Solution :

Price (X)	**dx = X–A$_x$ A$_x$ = 15**	**dx′ = $\frac{dx}{C_1}$ C$_1$ = 5**	**dx'^2**	**Demand (Y)**	**dy = Y–A$_y$ A$_y$ = 30**	**dx′ = $\frac{dy}{C_2}$ C$_2$ = 5**	**dy'^2**	**dx′ dy′**
5	– 10	– 2	4	40	10	2	4	– 4
10	– 5	– 1	1	35	5	1	1	– 1
15	0	0	0	30	0	0	0	0
20	5	1	1	25	– 5	– 1	1	– 1
25	10	2	4	20	– 10	– 2	4	– 4
N = 5		Σdx′ = 0	$\Sigma dx'^2$ = 10	N = 5		Σdy′ = 0	$\Sigma dy'^2$ = 10	Σdx′ dy′ = –10

$$r = \frac{\sum dx'dy' - \frac{(\sum dx')-(\sum dy')}{N}}{\sqrt{\sum dx'^2 - \frac{(\sum dx')^2}{N}} \times \sqrt{\sum dy'^2 - \frac{(\sum dy')^2}{N}}}$$

$$r = \frac{-10 - \frac{0}{5}}{\sqrt{10 - \frac{0}{5}} \times \sqrt{10 - \frac{0}{5}}}$$

$$r = \frac{-10}{\sqrt{10} \times \sqrt{10}} = \frac{-10}{10} = -1$$

Coefficient of Correlation (r) = –1.

This is a situation of a perfectly negative correlation between price and quantity demanded.

Properties of Correlation Coefficient

(1) r has no unit. It is a pure number. It means units of measurement are not parts of r.

(2) A negative value of r indicates an inverse relation, and if r is positive, the two variables move in the same direction.

(3) If r = 0, the two variables are uncorrelated. There is no linear relation between them. However, other types of relation may be there.

(4) If r = 1 or r = –1, the correlation is perfect or proportionate. A high value of r indicates strong linear relationship, i.e., +1 or –1.

(5) The value of the correlation coefficient lies between minus one and plus one, i.e., $-1 \le r \le +1$. If the value of r lies outside this range, it indicates error in calculation.

Spearman's Rank Correlation Coefficient

In 1904, **Charles Edward Spearman** developed a formula to calculate coefficient of correlation of qualitative variables. It is popularly known as *Spearman's Rank Difference Formula or Method*.

Attributes cannot be expressed in numbers or quantitative terms. Their relative merit can be determined on the basis of their order of preference or ranking.

It is in such situations that we use Spearman's Rank Difference Method.

Focus Formula

$$r_k = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Here, r_k = Coefficient of rank correlation;
D = Rank differences; and
N = Number of pairs.

Illustration : In a fancy-dress competition, two judges accorded following ranks to the 0 participants :

Judge X	1	2	3	4	5	6	7	8	9	10
Judge Y	10	6	5	4	7	9	8	2	1	3

Calculate coefficient of rank correlation.

Solution :

$$r_k = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Judge I, X = R_1	Judge II, X = R_2	D = R_1 – R_2	D^2
1	10	– 9	81
2	6	– 4	16
3	5	– 2	4
4	4	0	0
5	7	– 2	4
6	9	– 3	9
7	8	– 1	1
8	2	6	36
9	1	8	64
10	3	7	49
N = 10			$\Sigma D^2 = 264$

$$r_k = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Here, $N = 10$; $\Sigma D^2 = 264$

$$r_k = 1 - \frac{6 \times 264}{(10)^3 - 10} = 1 - \frac{1,584}{990}$$

$$= 1 - 1.6 = -0.6$$

Coefficient of Rank Correlation (r_k) = –0.6.

Coefficient of Rank Correlation when Ranks are Equal

Focus Formula

$$r_k = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + ...]}{N^3 - N}$$

Here, m = Number of items of equal ranks.

Illustration : Calculate coefficient of rank correlation between the marks in Economics and Statistics, as indicated by 8 answer books of each of the two examiners.

Marks in Statistics	15	10	20	28	12	10	16	18
Marks in Economics	16	14	10	12	11	15	18	12

Solution : There are 8 answer books each in Economics and Statistics indicating different marks. Rank 1 is accorded to the highest score. In Statistics, two answer books indicate 10 marks each. Hence, the first answer book has been given Rank 8 and the second 7. Thus, the average rank = $\frac{8+7}{2}$ = 7.5 has been accorded to both. Likewise, in Economics two answer books indicate 12 marks each. The average rank = $\frac{6+5}{2}$ = 5.5 has, therefore, been accorded to both.

Calculation of Coefficient of Rank Correlation

Marks in Statistics (X)	Rank R_1	Marks in Economics (Y)	Rank R_2	$D = R_1 - R_2$	D^2
15	5	16	2	3.0	9.00
10	7.5	14	4	3.5	12.25
20	2	10	8	– 6	36.00
28	1	12	5.5	– 4.5	20.25
12	6	11	7	– 1	1.00
10	7.5	15	3	4.5	20.25
16	4	18	1	3.0	9.00
18	3	12	5.5	– 2.5	6.25
N = 8					$\Sigma D^2 = 114$

Here, number 10 is repeated twice in series X and number 12 is repeated twice in series. Y. Therefore, in X, m = 2 and in Y, m = 2.

$$r_k = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)]}{N^3 - N}$$

$$= 1 - \frac{6[114 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{8^3 - 8}$$

$$= 1 - \frac{6[114 + \frac{1}{12}(6) + \frac{1}{12}(6)]}{512 - 8}$$

$$= 1 - \frac{6\left[114 + \frac{1}{2} + \frac{1}{2}\right]}{504} = 1 - \frac{6[115]}{504} = 1 - \frac{690}{504}$$

= 1 – 1.36 = –0.36

Coefficient of Rank Correlation (r_k) = – 0.36.

Pearson's method, popularly known as a Pearsonian Coefficient of Correlation, is the most extensively used quantitative methods in practice. The coefficient of correlation is denoted by "r".

If the relationship between two variables X and Y is to be ascertained, then the following formula is used :

Probable Error

Probable Error define the limits above and below size of the coefficient determined within which there is an equal chance that the coefficient of correlation similarly calculated form other sample will fall.

In other words, the probable error (P.E.) is the value which is added or subtracted from the coefficient of correlation (r) to get the upper limit and the lower limit respectively, within which the value of the correlation expectedly lies.

Focus Formula

$$P.E.r. = 0.6745\frac{1-r^2}{\sqrt{N}}$$

Where r = Coefficient of Correlation

N = Number of Observations

- There is no correlation between the variables if the value of 'r' is less than P.E. This shows that the coefficient of correlation is not at all significant.
- The correlation is said to be certain when the value of 'r' is six times more than the probable error; this shows that the value of 'r' is significant.
- By adding and subtracting the value of P.E from the value of 'r,' we get the upper limit and the lower limit, respectively within which the correlation of coefficient is expected to lie. Symbolically, it can be expressed
- ρ(rho) = r ± P.E.r.

Standard Error

Since Probable Error determine only 50% limit of Coefficient Correlation and it is not suitable in socio - economics and business areas, the modern statistics as adopted the use of standard errors in place of probable errors.

$$\boxed{S.E. = \frac{1-r^2}{\sqrt{N}}}$$

(3) Concurrent Deviation Method

This method of Studying correlation is the simplest of all the methods. When it is desired to study the direction of change rather than its quantity this method is suitable.

Focus Formula

$$r_C = \pm\sqrt{\pm\frac{(2c-n)}{n}}$$

where r_c = co-efficient of concurrent deviation.

c = number of concurrent deviations and

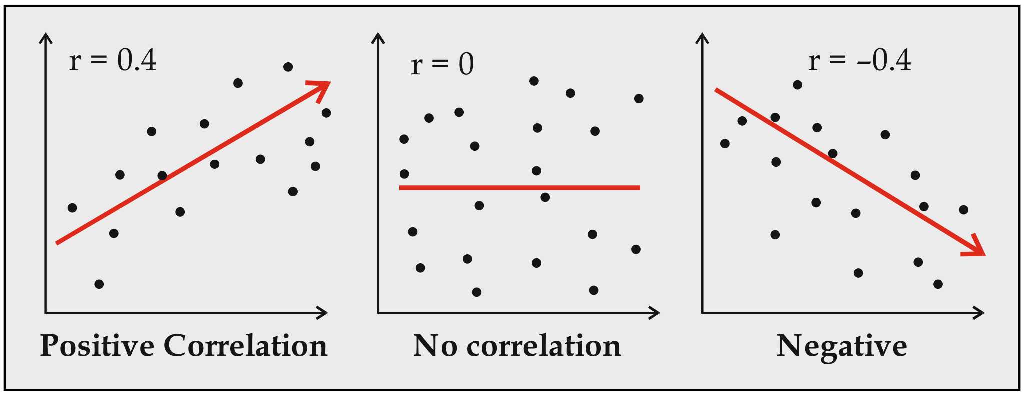
n = number of pairs of deviation.

Scatter of Plot

The first step is creating a scatter plot of the data. "There is no excuse for failing to plot and look. In general, scatter plots may reveal a

- Positive correlation (high values of X associated with high values of Y)
- Negative correlation (high values of X associated with low values of Y)
- No correlation (values of X are not at all predictive of values of Y).

These patterns are demonstrated in the figure to the right.



Ques. *Select the methods of finding out correlation from the following :*

(a) *Karl Pearson's Method* *(b)* *Spearman's Rank Method*

(c) *Yule's Method* *(d)* *Coefficient of Contingency*

(e) *Concurrent Deviation Method*

Codes : ***(NTA UGC-NET Dec. 2015 P-II)***

(1) *(a), (b), (c)* *(2)* *(a), (b), (c), (d)*

(3) *(a), (b), (e)* *(4)* *(c), (d), (e)*

Ans. *(3)* *These are methods of finding out correlation :*

(a) *Karl Pearson's Method*

(b) *Spearman's Rank Method*

(e) *Concurrent Deviation Method*

Ques. *The following are the estimated regression equations for x and y variables :*

$x = 0.85y$

$y = 0.89x$

With this information, the value of the coefficient of correlation would be :

(NTA UGC-NET Dec. 2015 P-III)

(1) *0.87* *(2)* *0.86*

(3) *0.89* *(4)* *0.75*

Ans. *(2)* *The following are the estimated regression equations for x and y variables :*

$x = 0.85y,\ y = 0.89x$

Coefficient of Correlation $= \sqrt{b_{xy}} \times b_{yx} = \sqrt{0.85} \times 0.89 = 0.869$

Ques. *Which one of the following formula is used to calculate probable error of correlation-coefficient between two variables of 'n' pairs of observations ?* ***(NTA UGC-NET June 2015 P-II)***

(A) $0.6745\frac{1-r^2}{\sqrt{n}}$ *(B)* $0.6745\frac{1-r^2}{\sqrt{n}}$

(C) $0.675\frac{1-r^2}{n}$ *(D)* $0.5758\frac{1-r^2}{n}$

Ans. *(A)* $0.6745\left[\frac{1-r^2}{\sqrt{n}}\right]$

Ques. *Which one of the following statements is false ?* ***(NTA UGC-NET June 2015 P-II)***

(A) Both correlation and regression co-efficients have same sign.

(B) Arithmetic mean of the regression co-efficients is always more than the correlation co-efficient.

(C) Regression co-efficients are independent of both the origin and scale.

(D) Correlation co-efficient is the square root of two regression co-efficients.

Ans. *(C) Regression co-efficients are independent of both the origin and scale. It is false.*

Que. *Match the items of List-I with the items of List-II and indicate the code of correct matching:*

	List-I		***List-II***
(a)	*Coefficient of determination*	*(i)*	$\gamma_{xy}\dfrac{\sigma_x}{\sigma_y}$
(b)	*Spearman's Rank correlation coefficient*	*(ii)*	$1-\dfrac{6\Sigma d^2}{n(n^2-1)}$
(c)	*Regression coefficient of x only y variable*	*(iii)*	$\dfrac{\Sigma xy}{n\sigma_x\sigma_y}$
(d)	*Karl Pearson's formula of calculating γ*	*(iv)*	γ^2

Codes : ***(NTA UGC-NET July 2016 P-II)***

	(a)	***(b)***	***(c)***	***(d)***
(1)	*(i)*	*(ii)*	*(iii)*	*(iv)*
(2)	*(i)*	*(iv)*	*(ii)*	*(iii)*
(3)	*(iv)*	*(iii)*	*(ii)*	*(i)*
(4)	*(iii)*	*(ii)*	*(iv)*	*(i)*

Ans. *(9) All options are wrong. Correct answer is :*

(a)	*(b)*	*(c)*	*(d)*
4	2	1	3

Ques. *Which one of the following formulae is used to calculate the standard error of coefficient of correlation between 25 paired observations of a sample ?* ***(NTA UGC-NET July 2016 P-III)***

(1) $\dfrac{(1-r^2)}{\sqrt{n}}$ *(2)* $\sqrt{\dfrac{(1-r^2)}{(n-2)}}$

(3) $(0.6745)\left(\dfrac{1-r^2}{\sqrt{n}}\right)$ *(4)* $\sqrt{\dfrac{(n-2)}{(1-r^2)}}$

Ans. *(2)* $\sqrt{\dfrac{(1-r^2)}{(n-2)}}$